AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 126, No. 6
Printed in U.S.A.

EXTERNAL COMPARISONS WITH THE CASE-COHORT DESIGN

SHOLOM WACHOLDER[1,2] AND JEAN-FRANÇOIS BOIVIN[1]

Wacholder, S. (Biostatistics Branch, NCI, NIH, Bethesda, MD 20892), and J.-F. Boivin. External comparisons with the case-cohort design. *Am J Epidemiol* 1987;126:1198–1209.

The case-cohort design can be an economic alternative to the standard cohort design. Prentice (*Biometrika* 1986;73:1–11) showed how the case-cohort design can be used to obtain relative risk estimates for comparisons within the cohort being studied. In this paper, the authors consider ways in which the case-cohort design can be used for comparing risk in exposure groups within the cohort to the risk in an external population. The problem reduces to estimating the number of expected cases at each exposure level in the total cohort, when exposure status is available only for members of a subcohort, i.e., a random sample of the total cohort. The authors describe theoretical and empirical properties of several variations of the design and analysis of case-cohort studies. Empirical properties were examined by replicating the selection of the subcohort in a study of second cancer risk after chemotherapy for a first cancer. Use of a case-cohort design in that study would have saved five-sixths of the cost of gathering covariate information at the price of only an 11% loss in efficiency relative to a full cohort study.

biometry; epidemiologic methods; follow-up studies; research design; statistics

Several questions in chronic disease epidemiology require the investigation of large populations in which the incidence of the disease of interest is low. Two examples are the association between exposure to high doses of ionizing radiation and incidence of leukemia, and the effect of prevention programs on coronary heart disease mortality. In a cohort study of leukemia risk after treatment for cervical cancer, Boice et al. (1) observed 77 leukemias among 82,616 women exposed to radiotherapy. In the Multiple Risk Factor Intervention Trial (2), 239 coronary heart disease deaths occurred among 12,866 men randomized to either a special intervention program or to their usual source of health care. These two investigations used standard epidemiologic designs, viz., the cohort study and the controlled prophylactic trial, respectively. In recent years, more economic approaches to the design and analysis of such investigations have been proposed by various authors. One example is the case-control-within-cohort or synthetic case-control design described by Liddell et al. (3). The case-cohort design described by Prentice (4) is another alternative.

In this paper, we concentrate on external as opposed to internal comparisons. In internal comparisons, the reference popula-

Received for publication April 28, 1986, and in final form January 6, 1987.

[1] Department of Epidemiology and Biostatistics, McGill University, Montréal, Québec, Canada.

[2] Current address: National Cancer Institute, Biostatistics Branch, Landow Building, Room 3C18, Bethesda, MD 20892. (Send reprint requests to Dr. Sholom Wacholder at this address.)

Supported in part by Public Health Service grant 2R01CA-22849 from the US National Cancer Institute and by a bourse d'été from l'Institut de recherche en santé et en sécurité du travail du Québec.

The authors thank their colleagues at Harvard University, McGill University, and the National Cancer Institute for their helpful comments on the manuscript; Celia Greenwood, University of Waterloo, for her help with computing; and Laurie Tesseris for assistance with the manuscript.

tion is a subset of the study cohort. By contrast, in external comparisons, a population not studied directly by the investigators is the reference group. In the first sections of this paper, we briefly discuss the standard cohort design, the case-control-within-cohort design, and the case-cohort design in the context of internal comparisons. We then focus on the case-cohort design and discuss its use in studies in which general population comparisons are required.

INTERNAL COMPARISONS

Full cohort studies

The classical approach to the investigation of large cohorts is the full cohort study, described in several standard textbooks of epidemiology (5–8). We use the word *cohort* to refer to a group of subjects who share a common characteristic and are followed after a specified point in time for the detection of new cases of disease. The common characteristic may be, for example, a blood group, employment at a particular plant, randomization into a clinical trial, or diagnosis of a specific disease. The actual calendar times of entry into the cohort often vary among members of the cohort. We restrict our attention to studies which require long follow-up, and where time since entry into the cohort is important, such as most studies of the incidence of cancer or of cardiovascular disease, as opposed to short follow-up studies such as the investigation of adverse pregnancy outcomes.

Figure 1 illustrates the full cohort study. A group of subjects is followed and the times of incidence of new cases of disease are ascertained. In the figure, each horizontal line represents a study subject. We assume, for concreteness, that there are 1,000 subjects in the total cohort, and that three of them developed the disease. In the full cohort design, all covariates, i.e., the exposures of interest and the potential confounders and effect modifiers, need to be measured and recorded for all members of the total cohort. The standard method of analysis for internal comparisons in full cohorts is the Cox proportional hazards regression model (9, 10), in which a risk set is formed at the time of occurrence of each case of disease. At each of these failure times, the vector of covariate values for the case is compared with those for the other subjects in the corresponding risk set. In the example in figure 1, three risk sets would be formed, one for each case of disease observed. Each risk set would include the case and all other subjects who survived at least until the point of diagnosis of disease in the case. The survivors to be included in the risk sets are those subjects who cross the dotted vertical lines in figure 1.

Case-control-within-cohort studies

Full cohort studies can be very expensive, especially when the number of subjects is very large. Two more economic options are the case-control-within-cohort design (3) and the case-cohort design (4).

Figure 2 illustrates how the study popu-

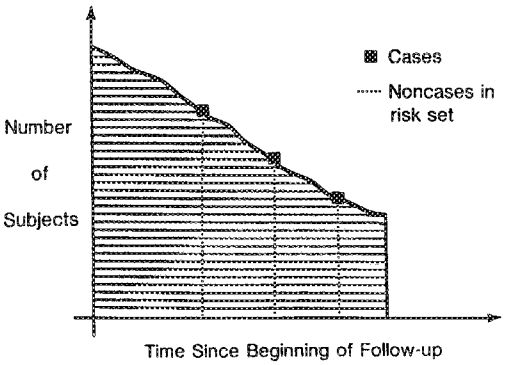

FIGURE 1. The full cohort design.

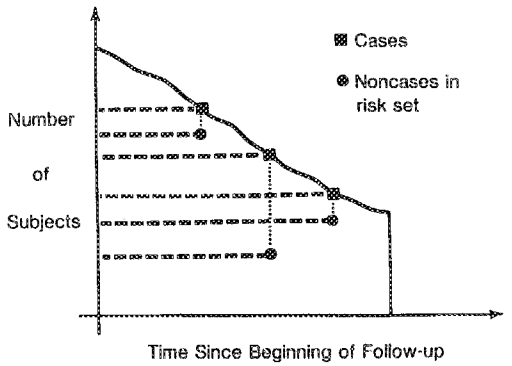

FIGURE 2. The case-control-within-cohort design.

lation used in the earlier example could be investigated with the use of the case-control-within-cohort design. Again, all 1,000 members of the total cohort must be followed. For each member of the total cohort, the dates of entry, termination of follow-up, and disease incidence, if any, would be ascertained. For each of the three cases, one or more controls are selected randomly from all subjects at risk at the time of occurrence of disease in the case. The standard analysis method is conditional logistic regression (11, 12). A moderate case to control ratio can achieve considerable cost savings without substantial loss of precision (11, 12).

Case-cohort studies

Figure 3 illustrates the use of the case-cohort design. Here, again, dates of entry, termination of follow-up, and disease incidence, if any, would be ascertained for all 1,000 members of the total cohort, and the same three cases would be observed. In the case-cohort design, however, a single random sample, or subcohort, would be obtained from all subjects in the total cohort. The three solid lines in figure 3 represent this sample.

The analysis of case-cohort studies for internal comparisons was discussed by Prentice (4). Risk sets are formed for each occurrence of disease. The noncases are those members of the subcohort at risk at the time of diagnosis of disease in each case.

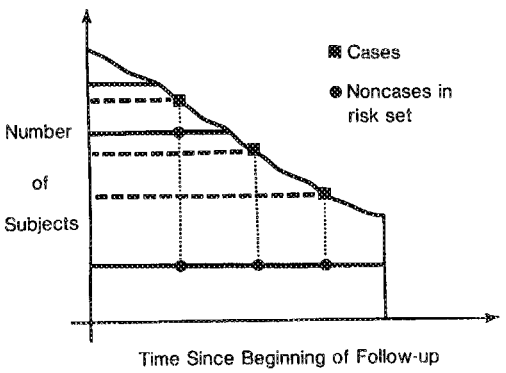

FIGURE 3. The case-cohort design.

Future cases in the subcohort *are* included as noncases; other cases *are not*. In our example in figure 3, the first risk set includes one case and two noncases, while the two other risk sets include one case and one noncase each. Relative risk estimates can be obtained by standard methods, though the calculation of standard errors is more complex (4). The case-cohort method is more economic than the full cohort study, and similar in economy to the case-control-within-cohort method, because exposures, confounders, and effect-modifiers must be measured only for the cases of disease and for the members of the subcohort.

While the sample chosen for the case-cohort study may be a simple random sample, a stratified random sample may be advantageous. Only variables which are ascertained for all members of the total cohort, such as year of birth, can be stratification variables. If, for example, more cases are expected to occur in subjects who are older at entry, the cohort could be sampled within age strata in order to increase the number of members of the subcohort who will be in the age strata in which failures are most likely to occur. In the internal analysis, the potential risk set for each case could then be restricted to those members of the subcohort in the same age stratum as the case, or to all members of the subcohort, with a regression adjustment for age (10). Selection probabilities can be set so that the distribution of expected cases into strata of the subcohort will be the same as the distribution of expected cases in the total cohort. This latter distribution, in turn, will be similar to that of the cases, assuming that the stratification variable does not modify the risk of disease for cohort members relative to the baseline risk in the population from which expected numbers are calculated.

EXTERNAL COMPARISONS

In the preceding paragraphs, we described the use of the full cohort design, the case-control-within-cohort design, and the case-cohort design in the context of inter-

nal comparisons. However, in some investigations, an external population for which high-quality incidence data are available can be a useful reference group (5, 6). In many cohorts, all subjects are exposed to some degree, though the dosage, time period, and agent may vary. Thus, an external population which is presumed to be almost entirely unexposed is an appropriate choice for estimation of risk due to a specified level of exposure relative to an exposure of zero. External comparisons, however, have their own limitations: most importantly, these comparisons may be relatively crude because measurements of covariates of importance may not be available for the reference population.

Data gathered from either the case-control-within-cohort design or the case-cohort design can be exploited to generate general population comparisons. Our objective in the following sections of this paper is to describe how the case-cohort design may be used for external comparisons and to assess the precision of the relative risk estimator from that design. The more complex problem of using data from a case-control-within-cohort study for external comparisons is the subject of ongoing research.

Full cohort studies

For external comparisons, the relative risk for subjects with a given exposure level is the ratio of the incidence rate for subjects in the cohort at that exposure level, relative to the incidence rate in a particular external population. The estimation of relative risk based on observed-to-expected ratios in full cohort studies is a routine procedure, extensively described in several textbooks of epidemiology (6). The numerator of the relative risk estimate $\widehat{\mathrm{RR}}_i$ for the ith exposure subgroup is

$$O_i = O_{i..} = \sum_d \sum_j O_{idj},$$

where the subscript d represents demographic covariates such as age, sex, and calendar year, and the subscript j represents the jth member of the cohort of size J. The dot subscript represents summation. The denominator is

$$E_i = E_{i..} = \sum_d \sum_j E_{idj} = \sum_d I_d \sum_j Y_{idj};$$

that is, the expected cases at the ith exposure level, which is the sum of the products of I_d, the incidence rate per unit time at demographic level d in the standard population, and the total person-time in the ith exposure group at demographic level d. Individuals may contribute to more than one d-specific person-time cell and each cell may contain the contribution of many members of the cohort. Dropping the j subscript for convenience, we obtain

$$\widehat{\mathrm{RR}}_i = \frac{O_i}{E_i} = \frac{O_{i.}}{E_{i.}} = \frac{\sum_d O_{id}}{\sum_d E_{id}} = \frac{\sum_d O_{id}}{\sum_d Y_{id} I_d}. \quad (1)$$

The circumflex above RR_i indicates an estimate based on the total cohort.

Case-cohort studies

In case-cohort studies, the observed cases are identified and studied exactly as in the full cohort study. However, the procedure used for calculation of E_i, the numbers of expected cases in full cohort studies, cannot be applied, because the required data on all members of the total cohort are not available. Nonetheless, case-cohort data can still be used to estimate expected numbers of cases in each exposure group. From equation 1,

$$E_i = \sum_d I_d Y_{id}. \quad (2)$$

To estimate E_i, we need the incidence rates I_d, known from general population statistics, and each of the i-d-specific sums of person-time Y_{id}, which are not directly available from the data. However, the amount of person-time for the *total* cohort can be estimated in the case-cohort design whenever the dates of entry and exit from the cohort are known for all study subjects.

The task is therefore to allocate the total person-time according to exposure status. Since the subcohort is a random sample from the total cohort, we can treat it as a "mini-cohort" from which person-time can be calculated for each exposure group. Using the lower case y's with the appropriate subscripts to refer to person-time calculated in the subcohort rather than in the total cohort, we can then obtain $y_{id}/y_{.d}$, the proportion of the person-time in any exposure category relative to the total amount of person-time in the subcohort. Then $Y_{.d}y_{id}/y_{.d}$ can be used to estimate Y_{id}, and the expected number in the ith exposure category can be estimated as:

$$\tilde{E}_{i,\mathrm{SR}} = \sum_d I_d \frac{Y_{.d}y_{id}}{y_{.d}}. \qquad (3)$$

The tilde above the E_i in equation 3 and in subsequent equations indicates values which depend on the particular subcohort that has been selected and which are used to estimate unknown values in the total cohort. Equation 3 can be rewritten in terms of expected numbers of cases:

$$\tilde{E}_{i,\mathrm{SR}} = \sum_d \frac{Y_{.d}I_d y_{id}I_d}{y_{.d}I_d} = \sum_d \frac{E_{.d}e_{id}}{e_{.d}}, \qquad (4)$$

where the lower case e's with the appropriate subscripts refer to expected numbers of cases calculated in the subcohort rather than in the total cohort.

Equations 3 and 4 describe a stratified analysis of case-cohort data. We refer to this method as the *stratified ratio* method, because the calculation of $\tilde{E}_i$ partitions the $E_{.d}$, obtained from the total cohort, in the ratio $e_{id}/e_{.d}$. However, when the subcohort from which the $e_{.d}$'s are calculated is a simple random sample of the total cohort from which the $E_{.d}$'s are obtained, the ratios $E_{.d}/e_{.d}$ are approximately constant over all values of d, and equation 4 can be further simplified to:

$$\tilde{E}_{i,\mathrm{UR}} = \frac{E_{.}}{e_{.}} \sum_d e_{id} = \frac{E_{.}e_i}{e_{.}}. \qquad (5)$$

Equation 5 gives a method of *unstratified ratio* analysis of case-cohort data, which can be used when the subcohort is a simple random sample of the total cohort. However, stratified random sampling can be used to increase the probability of selection of subjects from strata in which the baseline risk of disease is higher. This should lead to greater precision of the estimates of expected numbers of cases because the information used in the analysis then focuses more on subjects with larger contributions to the total expected number of cases rather than on subjects with lesser contributions. It also may increase the precision of internal estimates of relative risk, because, assuming that the stratification variable is not an effect modifier, the members of the subcohort and the cases will be distributed similarly into strata. When the subcohort has been selected with the use of stratified sampling rather than with simple random sampling, the ratios $E_{.d}/e_{.d}$ will generally not be constant over all d. In this case, the stratified analysis corresponding to equation 4 must be carried out.

The analyses described by equations 4 and 5 require knowledge of the vector d of demographic variables for all members of the total cohort in order to calculate the $E_{.d}$'s in equation 4, and $E_{.}$ in equation 5. However, in a case-cohort study, some demographic data may be collected only for the cases and members of the subcohort. Then, $E_{.}$ cannot be obtained directly; it can, however, be estimated unbiased as $e_{.}/f$, where f is the sampling fraction for selection of the subcohort from the total cohort. Equation 5 then becomes

$$\tilde{E}_{i,\mathrm{UX}} = \frac{e_{.}e_i}{e_{.}f} = \frac{e_i}{f}, \qquad (6)$$

which is unbiased for E_i. This simple method of estimation of E_i was used by Hutchison (13) and Boice and Hutchison (14). Smith and Doll (15) probably also used this approach for their dose-response analyses in their study of leukemia risk after radiotherapy for ankylosing spondylitis. We refer to this approach as the *unstratified expansion* method of analysis, and it can be used when the subcohort is a

simple random sample of the total cohort. In this method of analysis, the expected numbers derived from the total cohort are not used. The estimates of the E_i based on the expansion method are unbiased, while the ratio estimates of the E_i using equations 4 and 5 have an asymptotically negligible bias (16).

A more general method which allows different sampling fractions f_d within each demographic stratum is:

$$\tilde{E}_{i,\mathrm{SX}} = \sum_d \frac{e_{id}}{f_d}. \qquad (7)$$

We refer to this last method of analysis as the *stratified expansion* method.

Expressions for variance estimates, tests of hypotheses, and confidence intervals for external comparisons in case-cohort studies are developed in the Appendix.

Relative efficiency

It is possible to calculate the efficiency of the case-cohort method for estimation of the relative risk among members of the cohort with exposure i relative to the general population. In the full cohort design with a large cohort and relatively few failures, the random variable O_i, the observed number of cases at exposure level i, can be assumed to follow a Poisson distribution with mean and variance $\lambda_i E_i$, where λ_i is the relative risk parameter and E_i is the expected number of exposed cases. In the full cohort design, E_i can be treated as a constant. However, in the case-cohort design, E_i is not directly available and an estimate $\tilde{E}_i$ is obtained from the subcohort, adding variability to the relative risk estimate. The efficiency of the case-cohort design relative to the full cohort design for this exposure level is the inverse of the ratio of variances of the respective estimates of the relative risks when the relative risk is unity, or $E_i/[E_i + \mathrm{var}(\tilde{E}_i)]$.

Example

We investigated properties of the case-cohort design for general population comparisons in a cohort of 2,189 patients treated for Hodgkin's disease and followed for over 11,000 person-years. This cohort was studied to investigate the association between treatments for Hodgkin's disease and subsequent second cancer risk. Full treatment histories and the incidences of second cancers were obtained for all members of the total cohort. This study was analyzed using classical full cohort methods (6). Incidence data from the Connecticut Tumor Registry were used for external comparisons. Relative risk estimates for the internal and external comparisons are presented in an earlier paper (17), to which readers interested in the substantive results should refer. We reanalyzed these data with the use of the case-cohort design for external comparisons in order to determine the properties of the estimators of the expected number of second cancers after chemotherapy. Small discrepancies between the full cohort analyses reported here and in Boivin et al. (17) arose from minor differences in inclusion criteria, starting dates, and exposure definition.

Table 1 gives the distribution of study subjects and expected cancers in the total cohort, by age at diagnosis of Hodgkin's disease and year of diagnosis. All 2,189 study subjects were patients who survived at least one year after diagnosis of Hodgkin's disease. They were followed from the date of their one-year survival through 1978, death, loss to follow-up, or diagnosis of second cancer, whichever came first. For all 2,189 subjects, information was obtained on dates of entry and exit from the cohort; date of diagnosis of second cancer, if any; sex; and date of birth. A total of 72 cancers were observed in the total cohort, including 18 cancers after chemotherapy, while 29.22 cancers were expected in the total cohort on the basis of Connecticut general population incidence rates, including 3.57 cancers expected after chemotherapy.

For a case-cohort study, a subcohort was selected by obtaining a simple random sample of 288 subjects from the total cohort, for a case to subcohort ratio of 72:288 or 1:4. Information on exposure to chemo-

therapy was used only for these 288 subjects and the 72 cases of second cancer, for a maximum of 360 subjects for whom chemotherapy data were required, instead of 2,189 subjects in the full cohort study. In practice, chemotherapy data on slightly fewer than 360 subjects were necessary because of the overlap of cases and members of the subcohort. Person-years and expected cancers were then estimated for members of the subcohort exposed to chemotherapy, and for those not exposed. Results are given in table 2. The expected number in the entire subcohort was 3.385, including 0.544 after chemotherapy. These case-cohort data were then used to estimate the number of cancers expected after chemotherapy in the total cohort. The four methods of analysis described above were used: the stratified ratio method (SR), the unstratified ratio method (UR), the unstratified expansion method (UX), and the stratified expansion method (SX), corresponding, respectively, to equations 4–7.

To illustrate the use of the stratified ratio method of analysis corresponding to equation 4, we calculated expected numbers of cancers after chemotherapy within each of the six age and calendar year strata shown in tables 1 and 2. The percentages of expected cases after chemotherapy (exposure $= i = 1$) as a proportion of all expected cancers in each stratum of the subcohort are the ratios of the last two columns of table 2. Each ratio is then multiplied by the number of expected cancers in that stratum in the total cohort, shown in the last column of table 1, to obtain the estimate of the number expected after chemotherapy within each stratum in the total cohort. Then $\tilde{E}_{1,SR}$, the stratified ratio estimate of E_i, is the sum of these products:

$$\tilde{E}_{1,SR} = (0.000/0.303)3.09 + (0.129/0.736)5.14 + \cdots + (0.093/0.442)4.55 = 3.95.$$

TABLE 1

Study subjects and expected cancers in the total cohort, by age at diagnosis of Hodgkin's disease and year of diagnosis, Boston, Montreal, and New York patients, 1940–1975

Age (years) at diagnosis	Year of diagnosis	No. of subjects	Expected cancers after chemotherapy	All expected cancers
0–39	1940–1966	389	0.000	3.09
0–39	1967–1975	1,146	0.858	5.14
40–64	1940–1966	147	0.00067	5.37
40–64	1967–1975	397	2.24	10.25
≥65	1940–1966	22	0.000	0.84
≥65	1967–1975	88	0.479	4.55
Total		2,189	3.57	29.22

TABLE 2

Study subjects and expected cancers in a subcohort selected for a case-cohort study, by age at diagnosis of Hodgkin's disease and year of diagnosis, Boston, Montreal, and New York patients, 1940–1975

Age (years) at diagnosis	Year of diagnosis	No. of subjects	Expected cancers after chemotherapy	All expected cancers
0–39	1940–1966	46	0.000	0.303
0–39	1967–1975	158	0.129	0.736
40–64	1940–1966	14	0.000	0.276
40–64	1967–1975	53	0.322	1.580
≥65	1940–1966	2	0.000	0.048
≥65	1967–1975	15	0.093	0.442
Total		288	0.544	3.385

The unstratified ratio method of analysis, from equation 5, and using data given in tables 1 and 2, gives

$$\tilde{E}_{1,UR} = (29.22)(0.544)/3.385 = 4.70.$$

The unstratified expansion method, from equation 6, gives

$$\tilde{E}_{1,UX} = 0.544/(288/2{,}189) = 4.13,$$

while the stratified expansion method from equation 7 gives:

$$\tilde{E}_{1,SX} = 0.000/(46/389) + 0.129/(158/1{,}146) + \cdots + 0.093(15/88) = 3.89.$$

For the unstratified expansion method, the estimate of the relative risk in the exposed is

$$\tilde{RR}_1 = O_1/\tilde{E}_{1,UX} = 18/4.13 = 4.36.$$

With $s_1^2 = 0.000070$, equation A1 in the Appendix gives

$$\tilde{var}(\tilde{E}_{1,UX}) = (2{,}189^2)(1/288 - 1/2{,}189)(0.000070) = 1.01,$$

so that

$$\tilde{var}[\log(\tilde{RR})] = 1/18 + 1.01/4.13^2 = 0.115.$$

Thus, a 95 per cent confidence interval for $\tilde{RR}$ is

$$\exp\{\log 4.36 \pm 1.96\sqrt{0.115}\} \quad \text{or} \quad (2.24, 8.48).$$

The test statistic for the null hypothesis of a relative risk of unity is

$$\chi_1^2 = (18 - 4.13)^2/(4.13 + 1.01) = 37.4.$$

Precision of observed-to-expected ratios from the case-cohort design

The example shown in table 2 is not necessarily representative of the properties of the case-cohort design for this data set. We therefore replicated the selection of the subcohort and calculated the empirical bias and precision of the method. We obtained 300 samples of 144, 288, and 576 subjects, representing case to subcohort ratios of 1:2, 1:4, and 1:8, for two sampling schemes: simple random sampling, and random sampling stratified by age and calendar year, where the sampling fraction in each stratum was proportional to the number of expected cases in the stratum. The age and calendar year strata are shown in tables 1 and 2. For stratified random sampling, the sampling fraction in each stratum was proportional to the total number of expected cases in each stratum. In the stratified sampling scheme with a sample size of 144 subjects, for example, the number of subjects sampled among those who entered the cohort before age 40 years and before 1967 was (3.09/29.22) (144 subjects) = 16 subjects.

For each replication, we obtained four estimates of E_1, the expected number of cases after exposure to chemotherapy, obtained from studying the total cohort. We know E_1 for this cohort, but it will typically not be available when the case-cohort method is used. The relative sampling error is $\tilde{E}_1/E_1$, the ratio of the expected numbers of exposed subjects in the subcohort and in the total cohort. Since the number of cases observed among exposed is the same in the two designs, this relative sampling error reflects the additional variability introduced by use of the subcohort in estimation of the relative risk. We expected the average relative sampling error to be close to unity, but the spread of this error to vary according to subcohort size, sampling scheme, and analysis method. To estimate the spread, we obtained the fifth and 95th percentiles and the empirical variances of the relative sampling errors in our simulations. These can be compared to the average values of the variances obtained with the formulae in the Appendix. Table 3 shows results of these analyses. With simple random sampling in the selection of the subcohort from the total cohort, the four methods of analysis corresponding to equations 4–7 can be used. With stratified sampling, only stratified analyses are valid, as we showed in the derivation of equation 4.

TABLE 3

Average relative sampling error $\hat{E}_1/E_1$

Size of subcohort	Case to subcohort ratio	Sampling method*	Analysis method†	Average relative sampling error	5th and 95th percentiles for relative sampling error	Variance of replications	Average estimated variance
144	1:2	R	SR	1.017	0.473–1.758	0.143	0.108
		R	UR	0.990	0.474–1.632	0.136	0.129
		R	UX	0.978	0.451–1.624	0.135	0.127
		R	SX	0.968	0.455–1.581	0.124	0.108
		S	SR	1.028	0.634–1.461	0.0623	0.0644
		S	SX	1.018	0.601–1.450	0.0641	0.0676
288	1:4	R	SR	1.025	0.625–1.431	0.0690	0.0549
		R	UR	1.018	0.596–1.450	0.0638	0.0626
		R	UX	1.024	0.639–1.499	0.0683	0.0630
		R	SX	1.014	0.636–1.452	0.0601	0.0560
		S	SR	1.025	0.744–1.325	0.0351	0.0292
		S	SX	1.022	0.732–1.321	0.0312	0.0300
576	1:8	R	SR	1.023	0.753–1.314	0.0268	0.0246
		R	UR	1.015	0.758–1.317	0.0276	0.0268
		R	UX	1.022	0.764–1.319	0.0276	0.0266
		R	SX	1.018	0.761–1.295	0.0252	0.0242
		S‡	SR	1.015	0.858–1.175	0.00988	0.0105
		S‡	SX	1.015	0.863–1.173	0.00931	0.0107

* R is simple random sampling; S is stratified random sampling.

† SR is a stratified ratio analysis, UR is an unstratified ratio analysis, UX is an unstratified expansion analysis, and SX is a stratified expansion analysis, corresponding to equations 4–7, respectively, in the text.

‡ For the stratified sampling scheme, only 572 subjects were actually available because, with stratified sampling, our plan was that 16 per cent of the subjects in the sample would come from the category 65+ years of age, years of diagnosis 1967–1975, leading to a sample size of (0.16)(576) = 92 subjects for this category. However, the total cohort only included 88 subjects in this age-calendar year category.

We see in table 3 that, as we expected, average sampling errors are close to unity. There is no clear precision advantage to any of the analysis methods. On the other hand, a stratified sampling design does provide a substantial increase in precision. The averages of the variances using the formulae in the Appendix are near the empirical variances.

On the basis of data given in table 3, we calculated that a stratified sample with a case-to-subcohort ratio of 1:4 can result in 89 per cent efficiency since $\text{var}(\hat{E}_1) \cong E_1^2 \,\text{var}(\hat{E}_1/E_1) = (3.575^2)(0.0351)$. This loss of efficiency is compensated for by the economy of measuring covariates on only about one-sixth of the subjects in the total cohort.

DISCUSSION

The case-cohort design is not new. It has, for example, been used for a number of years for external comparisons in the field of cancer epidemiology (13–15). However, only recently has Prentice (4) presented a formal description of this design and showed how it could be used for internal comparisons. In our view, one advantage of the case-cohort design is, in contrast to the case-control-within-cohort design, the ability to carry out external comparisons simply. In case-control-within-cohort studies, the controls are not a random sample of the total cohort, thereby making external comparisons more complicated.

We reported little difference in the

performance of the expansion and ratio methods of estimation in the relatively small exposed group. Theoretical results in Cochran (16) suggest that the ratio method will tend to outperform the expansion method in exposure groups containing a large proportion of the subjects.

Stratifying in the selection of the subcohort on variables available for all members of the total cohort improved the performance of the case-cohort design for external comparisons, as we expect it would for internal comparisons. Sampling individuals disproportionately from strata in which the baseline risk of disease is higher gives more precision to external analyses than simple random sampling. For internal analyses, the individuals from strata with higher baseline risk will tend to contribute to more risk sets when the stratified proportional hazards model (9, 10) is used, because the number of failures should be proportional to the number of expected cases.

While we stressed one advantage of the case-cohort design over the case-control-within-cohort design, namely, the ability to carry out external comparisons simply, there are other factors which may favor the case-cohort design. One such advantage is that in studies in which the effect of a given exposure on disease risk is determined for several diseases, the same subcohort can be used repeatedly to estimate exposure-specific expected numbers of cases of each disease; in a case-control-within-cohort study, a new control series matched to the cases for the time variable is required for each case series. Of course, with the case-cohort design, the risk ratio estimates from studies of various diseases will be correlated, especially with subcohorts of small size.

A logistic advantage of the case-cohort design over the case-control approach is that it may be possible to select the subcohort at the outset of the study, when members of the total cohort are identified, without waiting for any processing of information about the case series.

There is a close relation among epidemiologic designs used to study cohorts (18). The full cohort design does not use any sampling. The case-cohort design samples randomly from the total cohort, independently of failure times. The case-control design samples referents at the failure times of the cases. The difference between the two economic designs is therefore simply whether referents are matched to cases on survival to time of failure of the case. In this sense, a case-cohort study may be seen simply as an unmatched case-control-within-cohort study, as described by Mantel (19).

We may then ask when should the case-control design be preferred to the case-cohort design for external comparisons. We conjecture that the case-cohort design is more efficient than the case-control design for external comparisons because the subcohort provides more complete information about the total cohort than does the set of controls selected at failure times only. Further work is required to compare the properties of these two designs for both internal and external comparisons.

REFERENCES

1. Boice JD Jr, Day NE, Anderson A, et al. Second cancers following radiation treatment for cervical cancer. An international collaboration among cancer registries. JNCI 1985;74:955–75.
2. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Intervention Trial. Risk factor changes and mortality results. JAMA 1982;248:1465–77.
3. Liddell FDK, McDonald JC, Thomas DC. Methods of cohort analysis appraisal by application to asbestos mining. J R Stat Soc (A) 1977;140:469–91.
4. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.
5. MacMahon B, Pugh TF. Epidemiology. Principles and methods. Boston: Little, Brown & Co., 1970.
6. Monson RR. Occupational epidemiology. Boca Raton, FL: CRC Press, Inc., 1980.
7. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research. Principles and quantitative methods. Belmont, CA: Lifetime Learning Publications, 1982.

8. Mausner JS, Kramer S. Mausner & Bahn epidemiology—an introductory text. Philadelphia: WB Saunders Co., 1985.
9. Cox DR. Regression models and life-tables. J R Stat Soc (B) 1972;34:187–202.
10. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: John Wiley and Sons, 1980.
11. Breslow NE, Day NE. Statistical methods in cancer research. Vol. 1. The analysis of case-control studies. IARC scientific publication no. 32. Lyon: IARC, 1980.
12. Breslow NE, Lubin JH, Langholz B, et al. Multiplicative models and cohort analysis. J Am Stat Assoc 1983;78:1–12.
13. Hutchison GB. Leukemia in patients with cancer of the cervix uteri treated with radiation. A report covering the first 5 years of an international study. JNCI 1968;40:951–82.
14. Boice JD, Hutchison GB. Leukemia in women following radiotherapy for cervical cancer: ten-year follow-up of an international study. JNCI 1980;65:115–29.
15. Smith PG, Doll R. Mortality among patients with ankylosing spondylitis after a single treatment course with x rays. Br Med J 1982;284:449–60.
16. Cochran WG. Sampling techniques. 2nd ed. New York: John Wiley and Sons, 1963:154–88.
17. Boivin JF, Hutchison GB, Lyden M, et al. Second primary cancers following treatment of Hodgkin's disease. JNCI 1984;72:233–41.
18. Miettinen OS. The "case-control" study: valid selection of subjects. J Chronic Dis 1985;38:543–8.
19. Mantel N. Synthetic retrospective studies and related topics. Biometrics 1973;29:479–86.
20. Rao CR. Linear statistical inference and its applications. New York: John Wiley and Sons, 1973.

Appendix

There is added variability in the estimation of the relative risk from case-cohort data due to the sampling used in obtaining $\hat{E}_i$, the case-cohort estimate of E_i, the expected number of cases among subjects at exposure level i. We now develop expressions for the variances of the four estimators $\hat{E}_i$ discussed in the article, and for the variance of the logarithm of the relative risk.

We regard the $E_{i,j}$, $j = 1, \ldots, J$, as a finite set of fixed constants, and the $e_{i,j}$, $j = 1, \ldots, n$, as a random sample of the $E_{i,j}$ selected without replacement. Since $\hat{E}_{i,\mathrm{UX}} = \frac{J}{n} e_{i.}$, finite sampling theory (16) can be used to obtain a variance estimate

$$\tilde{\mathrm{var}}(\hat{E}_{i,\mathrm{UX}}) = J^2\left(\frac{1}{n} - \frac{1}{J}\right)s_i^2, \tag{A1}$$

where

$$s_i^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} e_{i,j}^2 - e_{i.}^2/n\right].$$

By treating $E_.$, the number of expected cases in the total cohort, as a constant and using the delta method, or following Cochran (16), we can obtain an estimate of the variance of $\hat{E}_{i,\mathrm{UR}}$ as

$$\tilde{\mathrm{var}}(\hat{E}_{i,\mathrm{UR}}) = \tilde{\mathrm{var}}\left[\frac{E_. e_i}{e_.}\right] = E_.^2\tilde{\mathrm{var}}\left[\frac{e_i}{e_.}\right] = n^2\left(\frac{1}{n} - \frac{1}{J}\right)E_.^2\,\frac{e_{i.}^2}{e_.^2}\left[\frac{s_i^2}{e_{i.}^2} + \frac{s_.^2}{e_.^2} - \frac{2c}{e_{i.}e_.}\right] \tag{A2}$$

where

$$s_.^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} e_{.j}^2 - e_.^2/n\right]$$

and c, the covariance of $e_{i.}$ and $e_{.j}$, is

$$c = \frac{1}{n-1}\left[\sum_{j=1}^{n} e_{i.}e_{.j} - e_{i.}e_./n\right].$$

Variances for $\hat{E}_i$ from the stratified designs are obtained by summing stratum-specific variances.

It is convenient mathematically to address the variability of the logarithm of the estimate of the relative risk. For full cohort studies, dropping the i subscript, and using the Poisson assumption and the delta method (20), the asymptotic variance of $\log(\widehat{\mathrm{RR}}) = \log(O) - \log(E)$ is $1/(\lambda E)$. For the case-cohort design, $\log(\widetilde{\mathrm{RR}}) = \log(O) - \log(\tilde{E})$. Then, by the delta method and the fact that the selection of the subcohort is independent of disease status,

$$\tilde{\mathrm{var}}[\log(\widetilde{\mathrm{RR}})] = \tilde{\mathrm{var}}[\log(O)] + \tilde{\mathrm{var}}[\log(\tilde{E})] = \frac{1}{\lambda\tilde{E}} + \frac{\tilde{\mathrm{var}}(\tilde{E})}{\tilde{E}^2},$$

and

$$\widehat{\text{var}}(\widetilde{\text{RR}}) = \lambda^2 \widehat{\text{var}}[\log(\widetilde{\text{RR}})] = \lambda/\tilde{E} + \lambda^2 \widehat{\text{var}}(\tilde{E})/\tilde{E}^2.$$

The development above suggests an asymptotic one degree of freedom chi-square test of the null hypothesis that the relative risk is λ:

$$\chi_1^2 = [O - \lambda\tilde{E}]^2/[\tilde{E} + \lambda \text{ var}(\tilde{E})]$$

However, if the number of expected cases $\lambda\tilde{E}$ is small, a small-sample test is called for. Confidence limits can be obtained in the usual way from the estimate and variance of $\log(\widetilde{\text{RR}})$.